# Improving predictions of bacterial water quality with real-time networked sensors

Benjamen Wetherill

UMass Boston

School of Marine Sciences

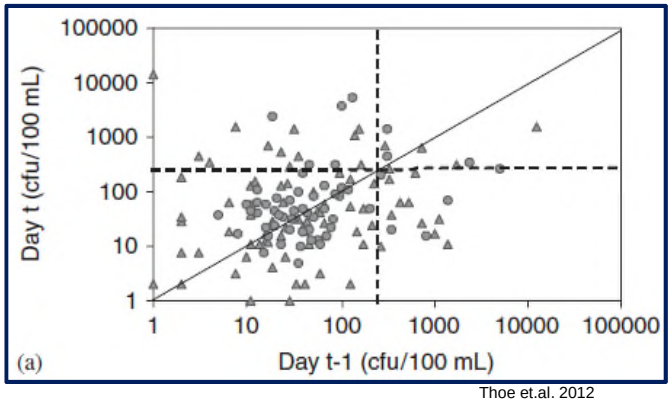
# Outline

- Methods for monitoring water quality
- Study sites & models developed
- Benefits of networked real-time sensors
- Results & Conclusions

# Current methods for monitoring water quality

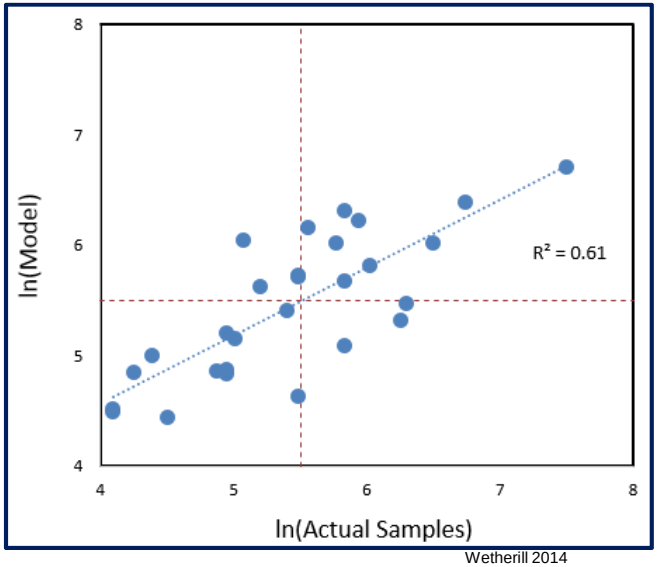
1. Lab analysis of water samples
   - Requires 24 hours
   - Handicapped by high variability of bacteria
   - Little correlation with next day conditions
2. Modeling based on environmental conditions
   - Positive correlation with rainfall, wind, turbidity ...
   - Negative correlation with solar radiance, salinity ...
   - 50-80% sensitivity

**Current vs. Previous Day**




Thoe et.al. 2012

**Model vs. Actuals**




Wetherill 2014

# Current practices employed by beaches

- Most beach managers use previous day sample results [1]
- Some combine that with simple rainfall thresholds
  - Mass DCR – sample positive or 24-hour rainfall > threshold
- Some use regression models based on hydro-meteorological data
  - CRWA – rain, temperature, wind, river flow
  - Ohio "Nowcasting" – turbidity, rainfall, lake stage, day of year
- Could not find any examples of real-time forecasts based on real-time data


Image credit: Amelia Kunhardt/The Patriot Ledger
http://www.patriotledger.com/x1351155712/11-South-Shore-beaches-closed-to-swimming

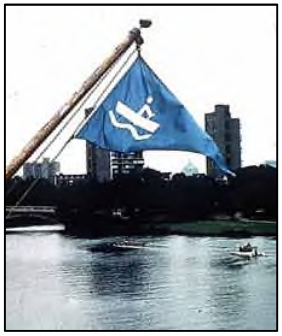


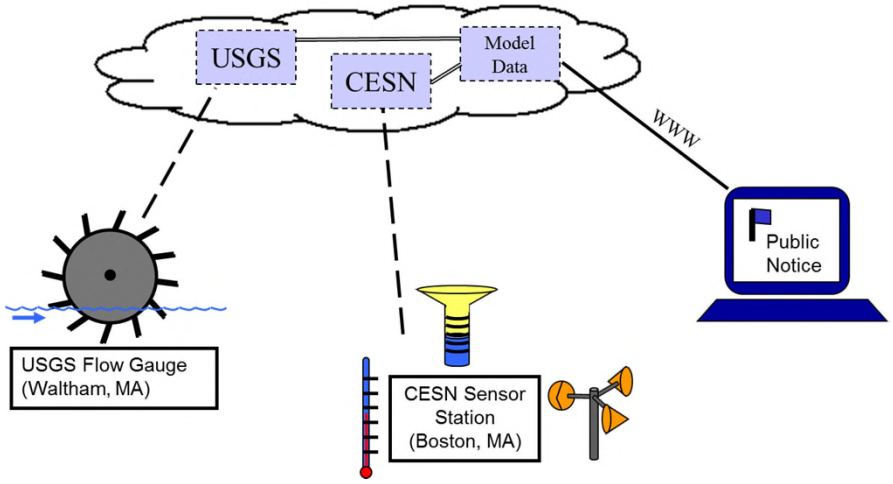
Image credit: CRWA
http://www.crwa.org/water_quality/daily/background.html


Image credit: Ohio Nowcast
http://www.ohionowcast.info/nowcast_how.asp

[ 1. NRDC 2004 ]

# Opportunity for Sensor Networks?

- What are sensor networks?




- With current technologies, it is completely feasible to automate data collection, analysis, and reporting.
- Could water quality modeling benefit from automated data?

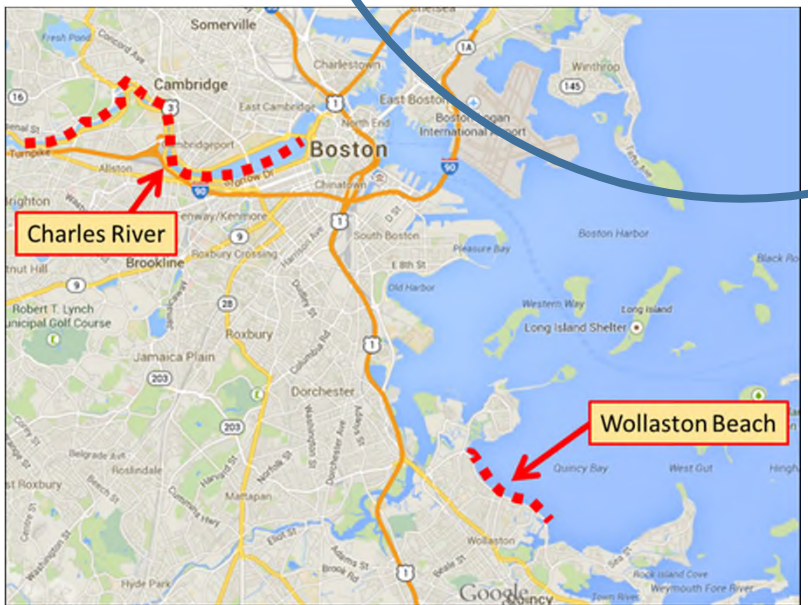
- Real-time data
  - Always updated with latest conditions
- Continuous data
  - Moving average trends instead of discrete points
- Localized data
  - Local data is always more relevant
- Connectable data
  - Data from multiple sources can be combined
- Automatable
  - Not dependent on a person being at work

# Outline

- Methods for monitoring water quality
- Sites studied & models developed
- Benefits of networked real-time sensors
- Results & Conclusions

# Two sites studied

- Charles River
  - Fresh water
  - River
  - *E.coli* bacteria
  - Samples 2x per week
  - 1½ seasons of data
  - Not open to swimming
- Wollaston Beach
  - Salt water
  - Bay
  - *Enterococcus* bacteria
  - Daily samples
  - 1 season of data
  - Open to swimming



# Wollaston Beach


Image credit: MWRA Environmental Quality Dept http://www.mwra.state.ma.us/harbor/graphic/wollaston.jpg

- Frequent beach closures
- Daily bacteria count data from DCR 6/25-8/30
- Environmental data sourced from...
  - CESN buoy
  - Boston tide gage
  - Umass weather station
- Modeled maximum daily bacteria from 4 sites

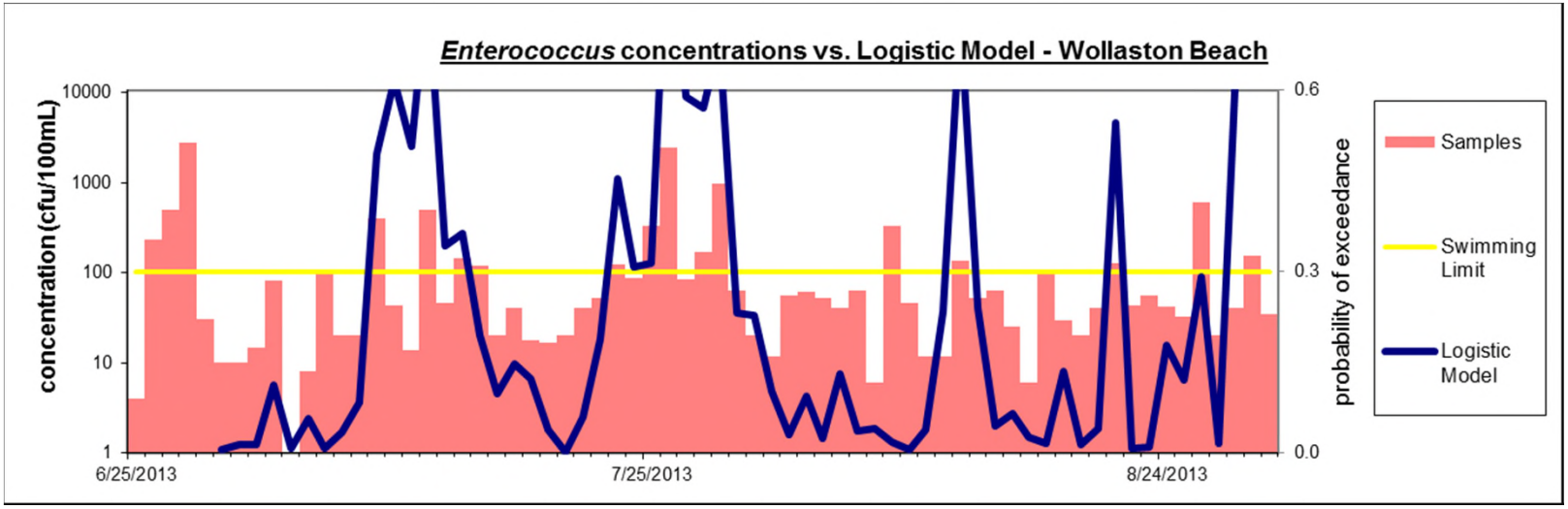
# Wollaston: Explanatory Variable Statistics

- Statistical analysis of each variable highlighted...
  - Wind direction
  - PAR
  - Air temp
  - Tide phase
  - Tidal range
  - Rain

| | | Corr. | $R^2$ | p-val |
|---|---|---|---|---|
| Wind Direction (1.5x cst.) | 24hr | 0.47 | 0.22 | 0.0001 |
| | 6hr | 0.44 | 0.20 | 0.0002 |
| PAR | 24hr | -0.37 | 0.14 | 0.0025 |
| | 12hr | -0.29 | 0.08 | 0.0213 |
| | 3hr | -0.25 | 0.06 | 0.0490 |
| Sample Light/Dark | | | 0.07 | 0.0382 |
| Air Temperature | 24hr | -0.38 | 0.15 | 0.0020 |
| Water Temperature | 24hr | -0.12 | 0.01 | 0.3588 |
| Sample Ebb/Flood | | | 0.12 | 0.0043 |
| Tidal Range | 48hr | 0.29 | 0.08 | 0.0199 |
| Sample Tide Level | | -0.08 | 0.01 | 0.5079 |
| Rain (log) | 72hr | 0.31 | 0.10 | 0.0111 |
| | 24hr | 0.31 | 0.09 | 0.0126 |
| Days since rain | | -0.23 | 0.05 | 0.0617 |
| Wind Speed (average) | 48hr | -0.15 | 0.02 | 0.2307 |
| | 24hr | -0.09 | 0.01 | 0.4697 |
| Salinity | 24hr | -0.06 | 0.00 | 0.6342 |
| Turbidity | 24hr | -0.02 | 0.00 | 0.8823 |
| Sample Time | | -0.16 | 0.03 | 0.1984 |

$$\ln\left(\frac{p}{1-p}\right) = 3.96 - 2.6T_p + 0.21\ln(R_{72} + 10^{-4}) - 0.40W_{24} + 0.22T_{24}M_{24}$$

# Wollaston: Model Performance (daily)




Linear Regression
True Positive: 50%; True Negative: 98%

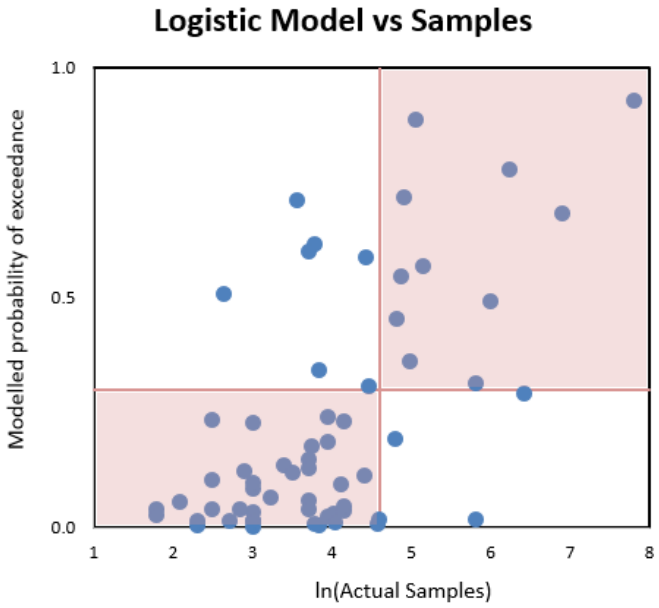
Logistic Regression (30% prob.)
True Positive: 79%; True Negative: 85%

Sensitivity

Specificity

Combined
True Positive: 86%; True Negative: 85%

Previous Day Bacteria
True Positive: 21% (2013), 14% (2012)

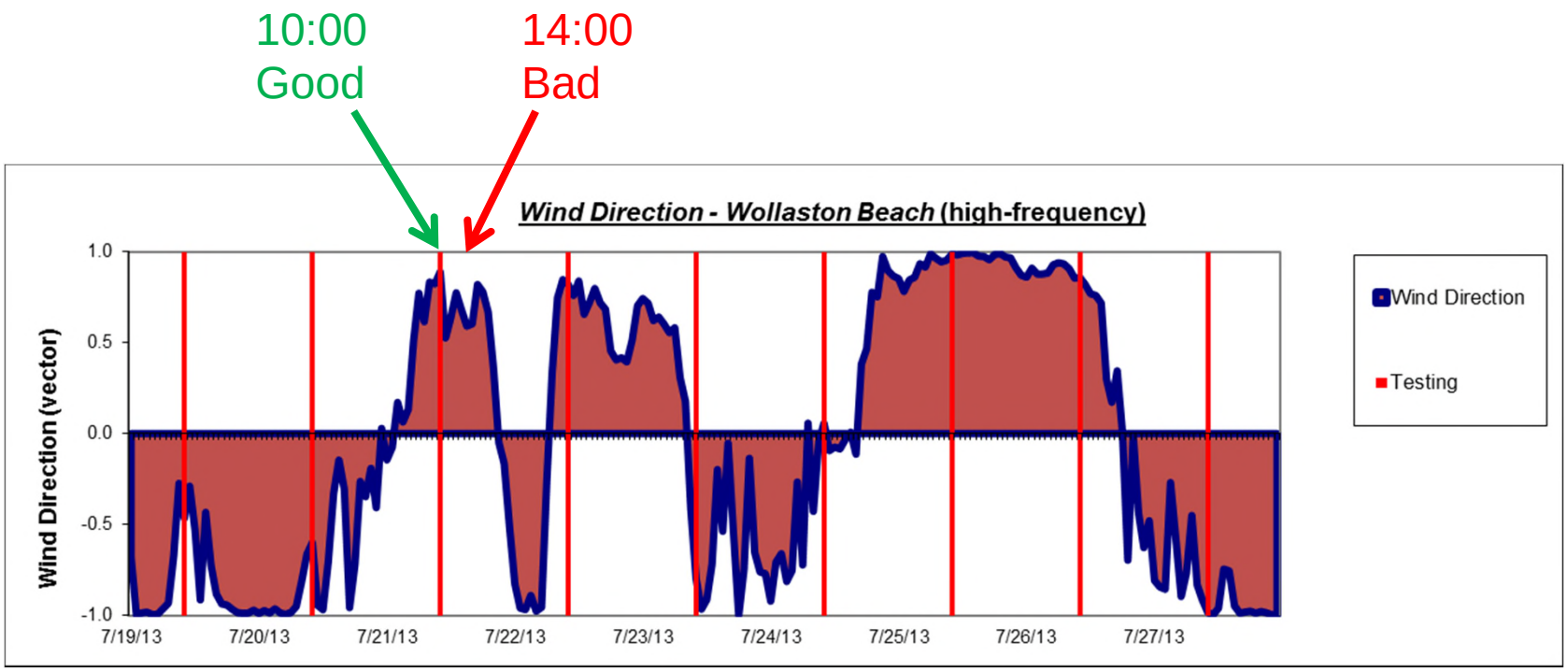


# Outline

- Methods for monitoring water quality
- Sites studied & models developed
- Benefits of networked real-time sensors
- Results & Conclusions

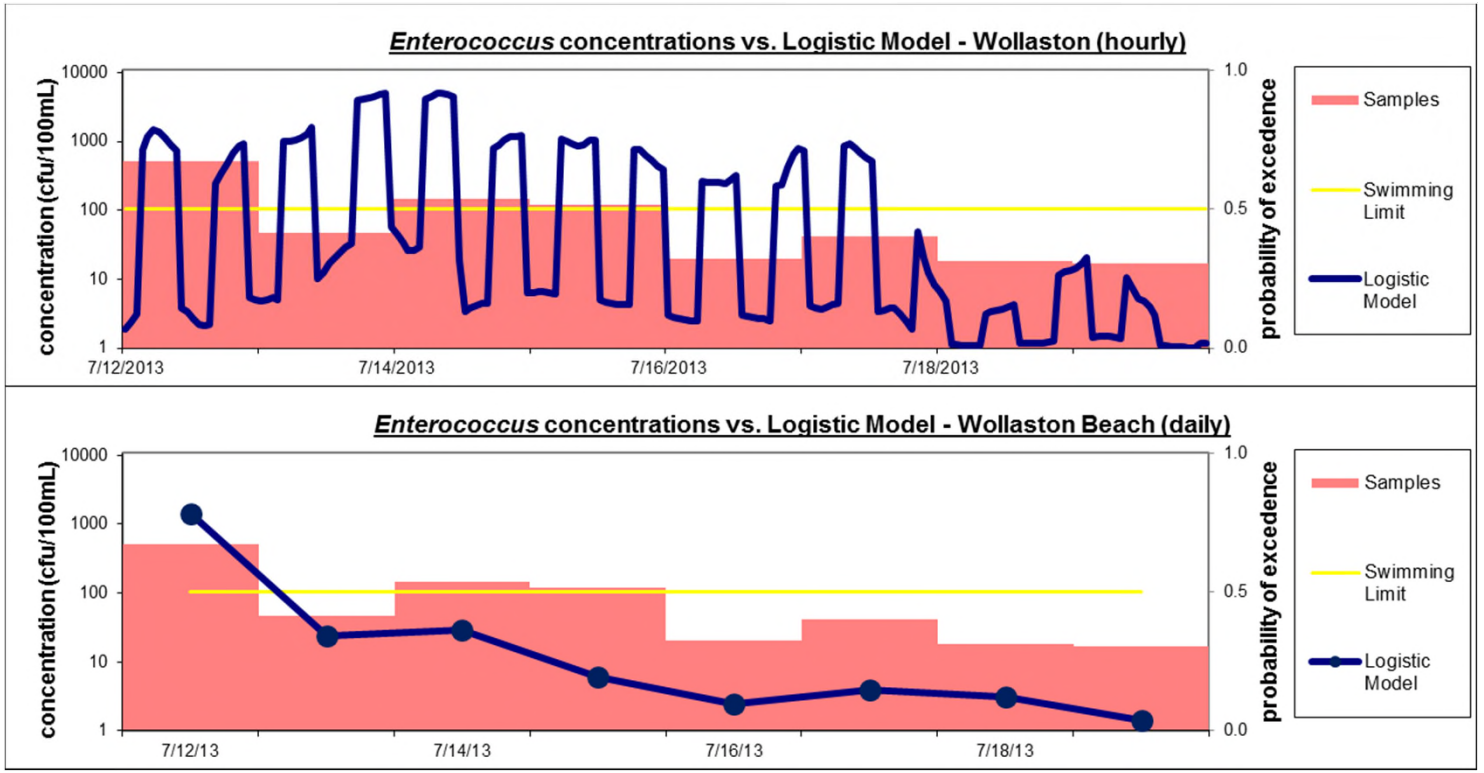
# Multiple updates per day

- Some conditions vary dramatically from hour to hour.
- Time of model calculation can significantly change results.



# Multiple updates per day

- Hourly model updates raised sensitivities above 90% at both sites ( > 5% increase).
  - Charles: daily – 87% TP ; hourly – 93% TP
  - Wollaston: daily – 79% TP ; hourly – 93% TP
- Frequent updates may also provide unknown information about non-sampled times of day.



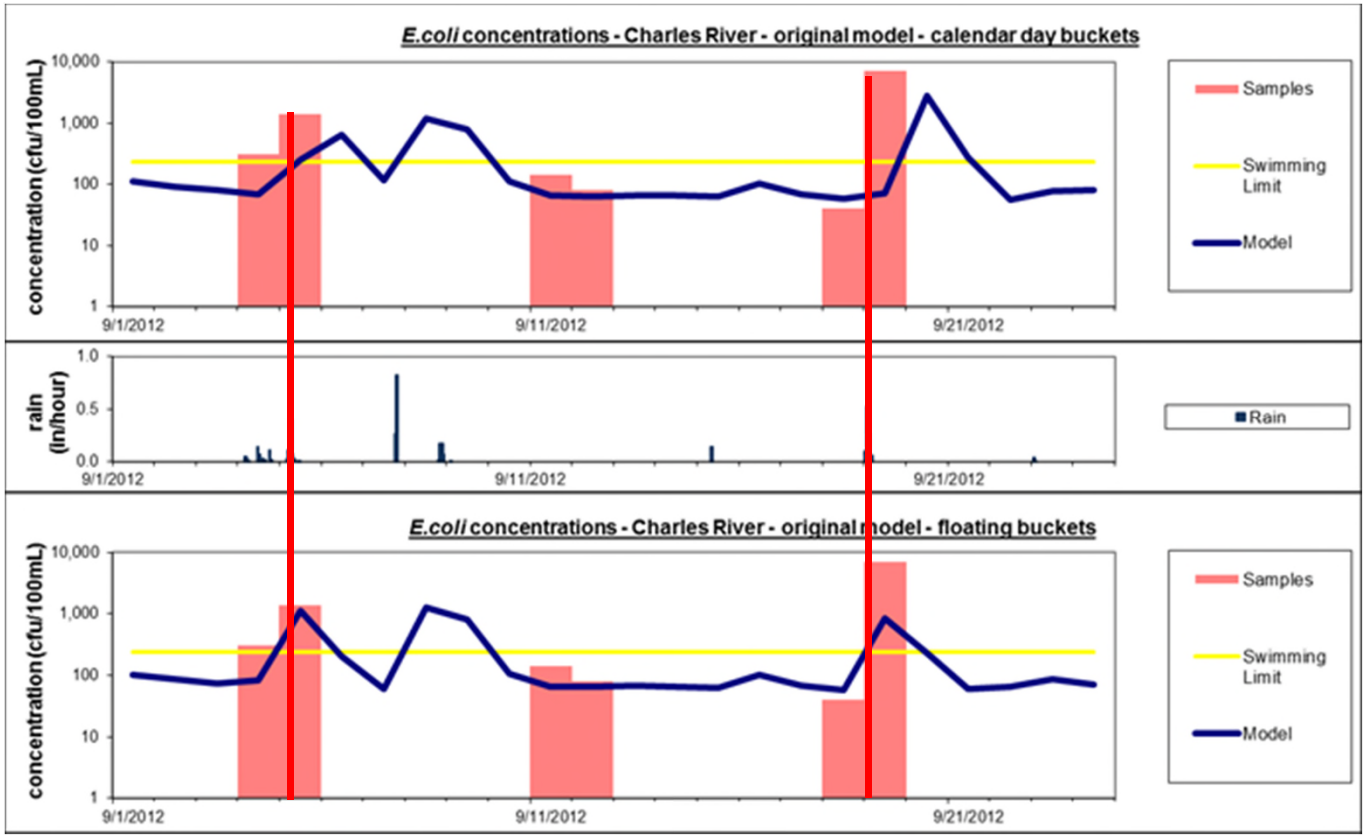
# Access to absolute latest data

- Ability to use latest data was key to Charles River model improvement.
  - Calendar day updates miss early morning events
  - Real-time updates capture early morning events

Old method
Calendar-day
updates

Rain

Real-time
updates



# Access to absolute latest data

- Can value of data timeliness be quantified?
- Testing adjusted time buckets resulted in > 5% loss of accuracy with only 4-hour old data

Real-time updates (aligned with sampling)

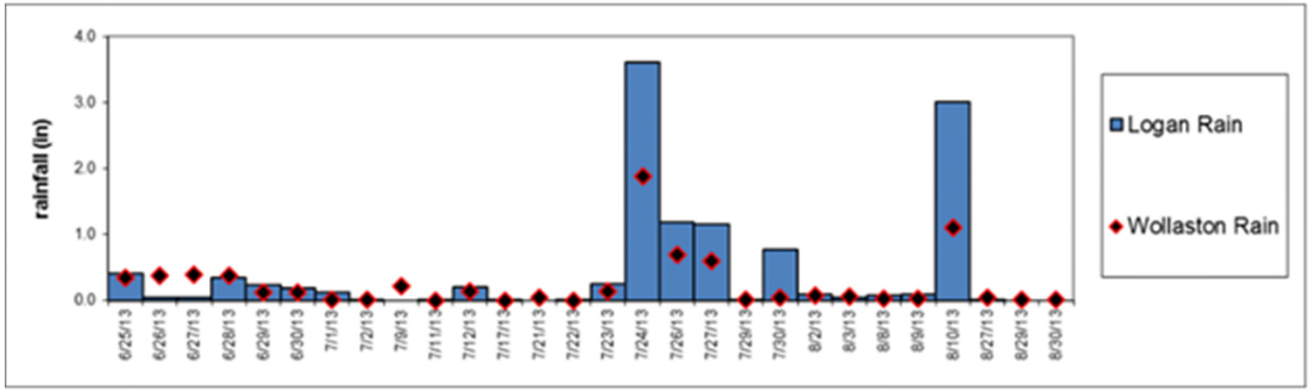
Median sampling time (7:33 AM)

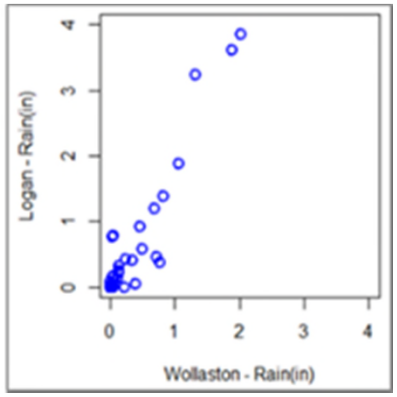
| Logistic Model | True positive | True negative | Accuracy |
|---|---|---|---|
| **Floating** | 79% | 85% | **84%** |
| **10:00 AM** | 64% | 70% | **69%** |
| **7:00 AM** | 79% | 74% | **75%** |
| **5:00 AM** | 86% | 70% | **74%** |
| **3:00 AM** | 64% | 70% | **69%** |

(Based on Wollaston 2013 data)

# Local data from local sensors

- Study evaluated benefit of local sensors
- Wollaston rain data collected from...
  - airport (6 mi) – larger big storms, smaller small storms
  - local station (3 mi) – consistently stronger correlations (~5% greater model accuracy)




| | Local Rain (UMass library) | | Logan Airport Rain | |
|---|---|---|---|---|
| Correlation with bacteria (24hr) | 0.31 | | 0.20 | |
| Correlation with bacteria (72hr) | 0.31 | | 0.21 | |
| | $R^2$ | p-value | $R^2$ | p-value |
| Regression predictive strength (24hr) | 0.09 | 0.013 | 0.04 | 0.107 |
| Regression predictive strength (72hr) | 0.10 | 0.011 | 0.04 | 0.089 |
| | Sensitivity | Accuracy | Sensitivity | Accuracy |
| Linear model | 50% | 87% | 43% | 82% |
| Linear model built with Logan rain | | | 43% | 80% |
| Logistic model | 79% | 84% | 71% | 79% |
| Logistic model built with Logan rain | | | 79% | 80% |

Some data can be sourced from existing sensors...

- River Flow – USGS gage 9 miles up river, but no tributaries in between
- Tide Height – NOAA gauge 6 miles away, but matches local tide within 10 minutes

# Outline

- Methods for monitoring water quality
- Sites studied & models developed
- Benefits of networked real-time sensors
- Results & Conclusions

# Results & Conclusions

- Modeling with environmental variables can provide significantly better predictions than previous day water samples
- Hourly updates provide higher sensitivity and useful data about sub-24-hour changes
  - >5% improvement in sensitivities over daily updates
- Real-time latest data results in more accurate models
  - >5% loss of accuracy when data is 4 hours old
- Some variables must be collected locally
  - 5% loss of accuracy when rain data sourced from airport instead of locally

# Acknowledgements


- Committee
  - Robert F. Chen, Crystal Schaaf, Bernie Gardner, Michael Shiaris
- Lab mates – Hayley Schiebel, Jill Arriola, Keith Cialino
- Francesco Peri - CESN
- Julie Wood – CRWA
- Kelly Coughlin – MWRA










# Questions?

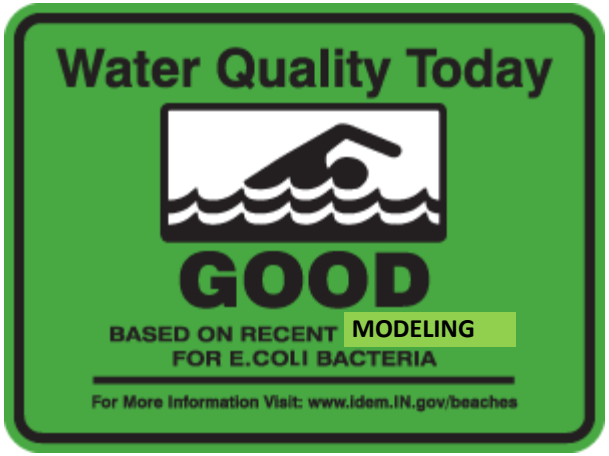



Image credit: Indiana Beaches http://www.in.gov/beachesalert/signs.html